COPERTINA

INDICE

1. Abstract
2. Linguaggio dei Segni
3. MediaPipe
4. Obiettivo (Risultato Atteso)
5. Implementazione
   1. StoreGesture.py
   2. Recogniser.py
6. Valutazione dei metodi implementati
7. Risultato Ottenuto
8. Condizioni di Luminosità
9. Requisiti Hardware and Software
10. Abstract

TODO

1. Linguaggio dei Segni

Le lingue dei segni sono lingue che veicolano i propri significati attraverso un sistema codificato di segni delle mani, espressioni del viso e movimenti del corpo. Sono utilizzate dalle comunità dei segnanti, a cui appartengono, in maggioranza, persone sorde.

La comunicazione attraverso la lingua dei segni avviene tramite il canale visivo-gestuale, producendo dei segni precisi, compiuti con una o entrambe le mani, aventi un significato specifico e assodato, come avviene per le parole.

I segni di ogni lingua dei segni possono essere scomposti in 4 componenti essenziali:

* Movimento,
* Orientamento,
* Configurazione,
* Luogo

(ossia le quattro componenti manuali del segno)

Le lingue dei segni sono lingue complete con una propria grammatica e un proprio dizionario, ma ne parliamo al plurale in quanto, ad ogni nazione, corrispondono diverse lingue dei segni; in Italia troviamo la **Lingua dei Segni Italiana** (**LIS**), negli Stati Uniti la **Lingua dei Segni Americana** (**ASL**), nel [Regno Unito](https://it.wikipedia.org/wiki/Regno_Unito) la [**Lingua dei Segni Britannica**](https://it.wikipedia.org/wiki/Lingua_dei_segni_britannica) (**BSL**), … e così via.

Motivo per cui è stato deciso di codificare una lista di segni "internazionali", che facilitasse il superamento delle [barriere](https://it.wikipedia.org/wiki/Barriera_linguistica) linguistiche.

Prende il nome di Signuno o Gestuno ed è la **Lingua dei Segni Internazionale**, sviluppata dalla World Federation of Deaf negli anni ’50, per permette la comunicazione tra persone sorde anche se di diversa nazionalità, e quindi usanti diverse lingue dei segni.

Trattandosi di una proposta esclusivamente lessicale, fu utilizzato, talvolta, in contesti di incontri internazionali, senza però acquisire mai le caratteristiche di una vera e propria lingua.

In Fig. 2.1. è possibile osservare i segni rappresentanti l’alfabeto del Signuno.



Fig. 2.1

1. MediaPipe

MediaPipe è una libreria multipiattaforma sviluppata da Google che fornisce incredibili soluzioni di Machine Learning pronte all'uso per attività di visione artificiale. Più specificatamente è un framework basato su grafi per la creazione di pipeline di machine learning multimodali (video, audio e sensori).

Tramite l’uso di Mediapipe e OpenCV è stato possibile implementare un prototipo scalabile di un Riconoscitore di gesti che abbiamo poi adattato al riconoscimento delle lettere dell’alfabeto della lingua dei segni internazionale.

* 1. MediaPipe Hands

Nonostante, per una persona, il riconoscimento di una mano e i suoi movimenti sia un task piuttosto semplice, ciò non è altrettanto corretto dal punto di vista di una macchina, infatti la percezione della mano in tempo reale è un compito di Computer Vision decisamente impegnativo, poiché le mani spesso si occludono a vicenda (ad es. occlusioni di dita/palmo, tremori e orientamento della mano) e ci si ritrova spesso in mancanza di pattern ad alto contrasto.

A tal proposito **MediaPipe Hands** è una soluzione di tracciamento di mani e dita ad alta fedeltà. Utilizza il Machine Learning (ML) per dedurre 21 punti di riferimento 3D di una mano da un solo fotogramma.

* + 1. Pipeline di Machine Learning

MediaPipe Hands utilizza una pipeline di Machine Learning composta da più modelli che lavorano insieme:

* Un **Palm Detection Model** che opera sull'immagine completa e restituisce un riquadro di delimitazione della mano orientato.
* Un **Hand Landmark Model** che opera sulla regione dell'immagine ritagliata definita dal rilevatore del palmo e restituisce punti chiave della mano 3D ad alta fedeltà.

Fornire l'immagine della mano accuratamente ritagliata all’**Hand Landmark Model** riduce drasticamente la necessità di aumentare i dati (ad esempio rotazioni, traslazione e scala) e consente invece alla rete di dedicare la maggior parte della sua capacità all'accuratezza della previsione delle coordinate.

* + 1. Palm Detection Model

La mancanza di motivi ad alto contrasto nelle mani rende relativamente difficile rilevarle in modo affidabile solo dalle loro caratteristiche visive. Invece, fornire un contesto aggiuntivo, come le caratteristiche del braccio, del corpo o della persona, aiuta la localizzazione accurata della mano.

Il **Palm Detection Model** nasce dal fatto che stimare i riquadri di delimitazione di oggetti rigidi come palmi e pugni è significativamente più semplice rispetto al rilevamento di mani con dita articolate. Inoltre, i palmi possono essere modellati utilizzando riquadri di delimitazione quadrati (“*ancore*” nella terminologia ML) ignorando altre proporzioni e quindi riducendo il numero di ancoraggi di un fattore 3-5. In secondo luogo, un estrattore di funzionalità codificatore-decodificatore viene utilizzato per una maggiore consapevolezza del contesto della scena anche per piccoli oggetti. Infine, la perdita focale viene ridotta al minimo durante l'allenamento per supportare una grande quantità di ancoraggi risultanti dalla varianza su larga scala.

Con le tecniche di cui sopra, si ottiene una precisione media del 95,7% nel rilevamento del palmo.

* + 1. Hand Landmark Model

Dopo il rilevamento del palmo sull'intera immagine, il successivo **Hand Landmark Model** esegue una precisa localizzazione dei punti chiave di 21 coordinate 3D delle nocche della mano all'interno delle regioni della mano rilevate tramite regressione, ovvero la previsione diretta delle coordinate. Il modello apprende una rappresentazione coerente della posa della mano interna ed è robusto anche per mani parzialmente visibili e in presenza di auto-occlusioni.

Per ottenere il risultato sono state annotate manualmente ~30.000 immagini del mondo reale con 21 coordinate 3D, come mostrato in Fig.1.1 (il valore Z viene preso dalla mappa di profondità dell'immagine, se esiste, per ogni coordinata corrispondente). Per coprire meglio le possibili pose della mano e fornire un'ulteriore supervisione sulla natura della geometria della mano, viene eseguito anche il rendering di un modello di mano sintetico di alta qualità su vari sfondi e lo mappiamo alle corrispondenti coordinate 3D.

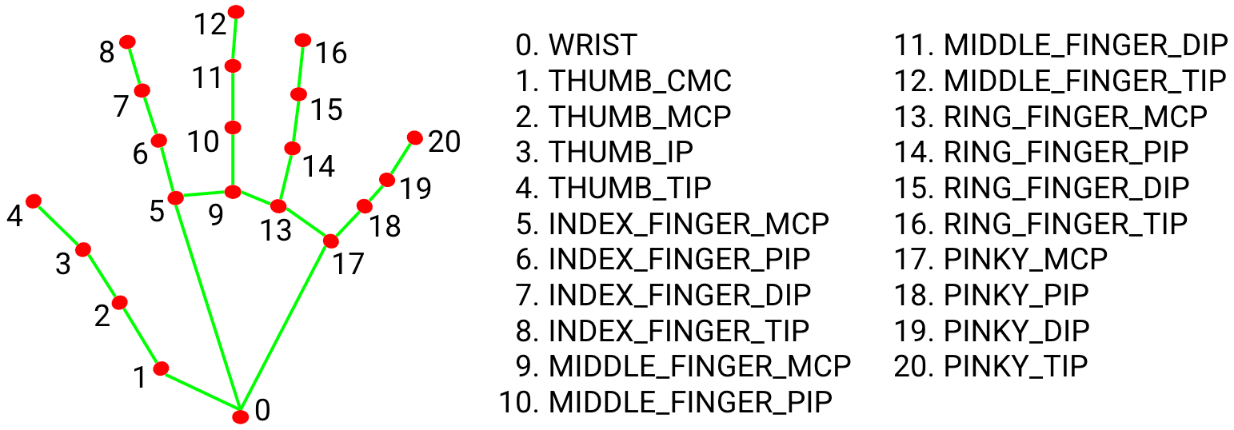


Fig. 3.1

* + 1. Configuration Options
* **STATIC\_IMAGE\_MODE**: Se impostato su ***false***, la soluzione tratta le immagini di input come un flusso video. Tenterà di rilevare le mani nelle prime immagini di input e, in caso di rilevamento riuscito, localizzerà ulteriormente i punti di riferimento della mano. Nelle immagini successive, una volta rilevate tutte le mani e localizzati i corrispondenti punti di riferimento delle mani, tiene semplicemente traccia di tali punti di riferimento senza invocare un altro rilevamento fino a quando non perde traccia di tutte le mani. Ciò riduce la latenza ed è ideale per l'elaborazione di fotogrammi video. Se impostato su ***true***, il rilevamento delle mani viene eseguito su ogni immagine di input, ideale per l'elaborazione di un batch di immagini statiche, possibilmente non correlate. Il valore di default è *false*.
* **MAX\_NUM\_HANDS**: Il numero massimo da mani da rilevare. Il valore di default è 2.
* **MODEL\_COMPLEXITY**: Complessità dell’hand landmark model. Il valore può essere **0** o **1**. Con l’aumento della complessità del modello, l’accuratezza del punto di riferimento e la latenza dell’interferenza, di solito, aumentano. Il valore di default è 1.
* **MIN\_DETECTION\_CONFIDENCE**: Rappresenta la misura di quanto il rilevamento della mano sia considerato riuscito da parte del modello. È un valore compreso nell’intervallo **[0.0, 1.0]**. Il valore di default è 0.5.
* **MIN\_TRACKING\_CONFIDENCE**: Valore di confidenza minimo (nell’intervallo **[0.0, 1.0])** richiesto dal modello di tracciamento dei punti di riferimento affinché questi siano considerati tracciati correttamente, altrimenti il ​​rilevamento della mano verrà richiamato automaticamente sull'immagine di input successiva. Impostandolo su un valore più elevato è possibile aumentare la robustezza della soluzione, a scapito di una latenza più elevata. Questo valore è ignorato se STATIC\_IMAGE\_MODE è settato a *True*, dove il rilevamento delle mani viene eseguito semplicemente su ogni immagine. Il valore predefinito è 0.5.
  + 1. Output
* **MULTI\_HAND\_LANDMARKS**: Raccolta di mani rilevate/tracciate, in cui ogni mano è rappresentata come un elenco di 21 punti di riferimento, ciascuno di essi composto da 3 coordinate **(x, y, z)**. ‘x’ e ‘y’ sono normalizzati a **[0.0, 1.0]** rispettivamente per la larghezza e l'altezza dell'immagine. ‘z’ rappresenta la profondità del punto di riferimento con la profondità al polso come origine: minore è il valore più il punto di riferimento è vicino alla fotocamera. La grandezza di z utilizza, all'incirca, la stessa scala di ‘x’.
* **MULTI\_HAND\_WORLD\_LANDMARKS**: Raccolta di mani rilevate/tracciate, dove ciascuna mano è rappresentata come un elenco di 21 punti di riferimento nelle coordinate del mondo. Ogni punto di riferimento è composto da tre coordinate (x, y e z): coordinate 3D (in metri) del mondo reale con l'origine nel centro geometrico approssimativo della mano.
* **MULTI\_HANDEDNESS**: Raccolta della manualità delle mani rilevate/tracciate (cioè si tratta di una mano sinistra o destra?). Ogni mano è composta da un’etichetta e un punteggio. “**label**” è una stringa di valore "***Left***" o "***Right***". “**score**” è la probabilità stimata della manualità prevista ed è sempre 0.5 (e la manualità opposta ha una probabilità stimata di 1 - score).

Si noti che la manualità è determinata assumendo che l'immagine in ingresso sia speculare, cioè scattata con una fotocamera frontale/selfie con immagini capovolte orizzontalmente. In caso contrario, è necessario scambiare l'output della manualità nell'applicazione. (Nel nostro caso abbiamo, infatti, flippato l’immagine)

1. Obiettivo

L’obiettivo che ci siamo prefissati di raggiungere in questa tesina consiste nella realizzazione di un **assistente visivo virtuale** che, mediante l’utilizzo di una webcam, sarà in grado di riconoscere e tradurre i segni riprodotti dalla mano sinistra, in real-time.

Tale assistente visivo virtuale sarà un programma OpenCV, con utilizzo del Framework Mediapipe per migliorare il modello di riconoscimento della mano, capace di tradurre i gesti riprodotti dalla mano in lettere dell’alfabeto della lingua dei segni internazionale (Fig 2.1).

Questo progetto è il primo passo nella creazione di un potenziale traduttore delle lingue dei segni, tale da poter riconoscere la comunicazione in lingua dei segni e tradurla in lingua scritta istantaneamente.

Un tale traduttore potrebbe ridurre notevolmente la barriera tra molte persone non udenti e ipoudenti in modo che possano migliorare la comunicazione con gli altri nelle loro attività quotidiane.

1. Implementazione
   1. HandTrackingModule.py
      1. Configurazione di MediaPipe

Nella nostra applicazione il modello Mediapipe è stato configurato con i valori in tabella.

|  |  |  |
| --- | --- | --- |
| **Opzioni di Configurazione** | **Valore** | **Intervallo di Riferimento** |
| STATIC\_IMAGE\_MODE | **False** | True/False |
| MAX\_NUM\_HANDS | **1** | 1, 2, …, n unsigned int |
| MODEL\_COMPLEXITY | **1** | 0/1 |
| MIN\_DETECTION\_CONFIDENCE | **0.5** | [0.0, 1.0] |
| MIN\_TRACKING\_CONFIDENCE | **0.6** | [0.0, 1.0] |

* + 1. Rilevamento della mano sinistra

Per i settaggi scelti su MediaPipe, il nostro sistema sarà in grado di individuare una mano alla volta, quindi se all’interno dell’inquadratura saranno visibili entrambe le mani o più mani, verrà riconosciuta la prima mano che entrerà nel campo visivo.

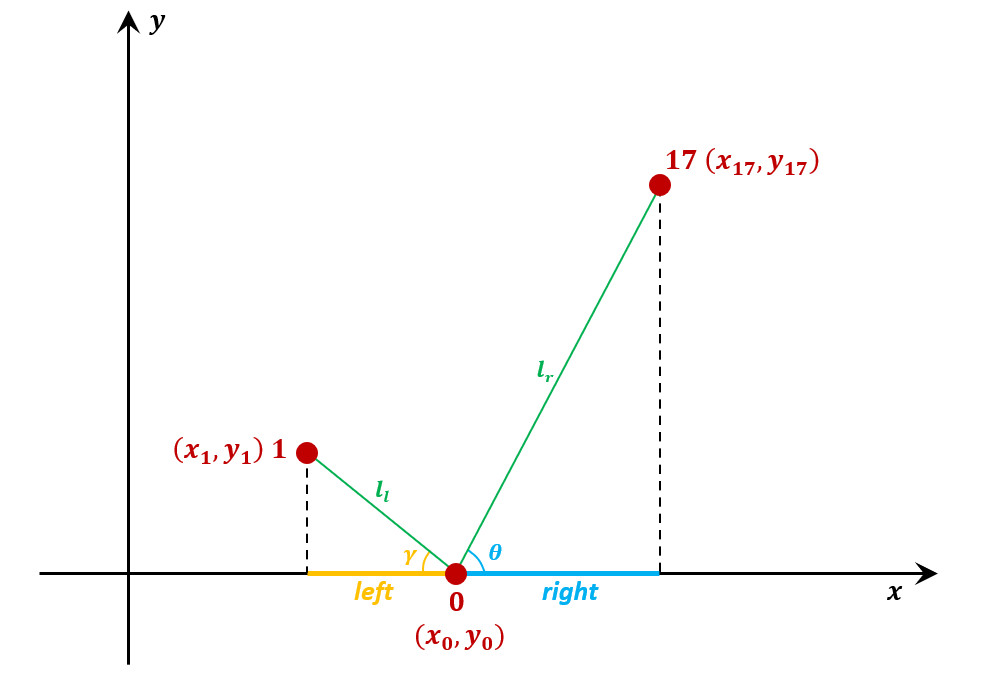
Immagine che contiene testo

Descrizione generata automaticamenteLa funzione che si occupa di rilevare la mano è **find\_hands**. La quale, tramite *MULTI\_HAND\_*LANDMARK, che è la raccolta di mani rilevate/tracciate, e *MULTI\_HANDEDNESS,* che è la raccolta della manualità delle mani rilevate/tracciate, (viste più nel dettaglio nel paragrafo 3.1.5), sarà in grado di riconoscere la mano una volta entrata nell’inquadratura e sapere distinguere tra destra e sinistra. Nel caso in cui venisse rilevata la mano sinistra, questa funzione ritornerà all’utente un feedback visivo del tracciamento della mano.

La funzione **right\_hand** invece, similmente al meccanismo della funzione precedente, ritornerà un errore nel caso di rilevamento della mano destra.

* + 1. Find\_position() **ROBERTA**
    2. Orientamento

Per semplicità consideriamo il seguente sistema di riferimento, rappresentando i nodi 0, 1 e 17 della mano, l’asse x è coerente col modello utilizzato da MediaPipe, mentre l’asse y no, in quanto la massima coordinata y, in MediaPipe, si raggiunge in corrispondenza del punto 0. Tuttavia, ai fini del calcolo dell’orientamento, questa differenza è ininfluente.



Per il Teorema di Pitagora abbiamo che:

Quindi si ha che:

Sostituendo:

Il coseno al quadrato è una funzione limitata nell’intervallo per cui basta trovare il giusto intervallo entro cui la mano risulta orientata più o meno verticalmente.

L’intervallo di è più ampio in quanto la distanza che intercorre tra i nodi 0 e 1 è minore rispetto a quella fra i nodi 0 e 17.

* 1. Recogniser.py

**ROBERTA**

* 1. RecogniserSigned.py

**ROBERTA**

* 1. StoreGestures.py

**ROBERTA**

* 1. StoreGestureSigned.py

**ROBERTA**

1. Valutazione dei metodi implementati

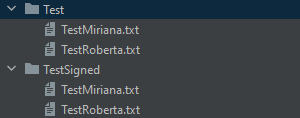
Nel capitolo precedente abbiamo illustrato i due approcci che abbiamo usato per lo sviluppo del **Recogniser.py** e del **RecogniserSigned.py**.

Ritenendo entrambi gli approcci validi e con prestazioni più o meno equivalenti, in quanto il primo riconosce i gesti con più facilità ma ha meno accuratezza mentre il secondo li riconosce con meno facilità ma in modo più distintivo, abbiamo sviluppato dei test per poter stabilire in modo oggettivo quale dei due approcci portasse risultati migliori.

Il test che abbiamo effettuato sui due metodi è molto semplice in quanto va a vedere per ogni lettera quante volte, su un campione di 100 frame, il programma riconosce la lettera dato il gesto corrispettivo, in poche parole determina i true positive.

Immagine che contiene testo

Descrizione generata automaticamenteQuesto pezzo di codice, che spiega come abbiamo calcolato i true positive, è estratto da *Test.py* (equivalente in *TestSigned.py*) all’interno della funzione **test:**

Immagine che contiene testo

Descrizione generata automaticamenteDopo aver determinato, a testa, la percentuale di riuscita di tutte le lettere dell’alfabeto, per entrambi gli approcci, ed averli salvati in dei file di questa tipologia:

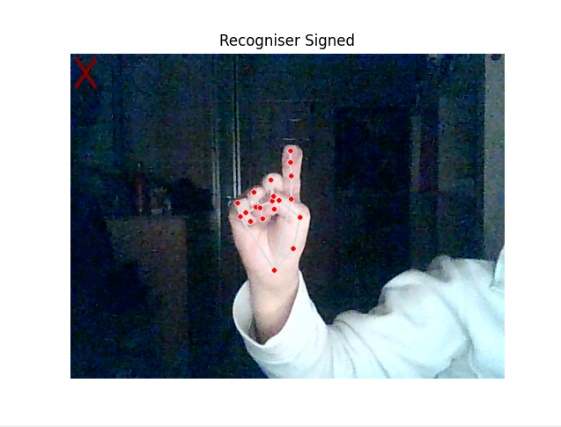
Abbiamo calcolato la riuscita media, tramite l’*AverageCalculator.py,* ottenendo una percentuale di riuscita del **Recogniser.py** del ***90.99%***e del **RecogniserSigned.py** del ***92.04%***.

In conclusione, prendendo in considerazione solo i test compiuti da noi, possiamo quindi evincere che entrambi i metodi sono abbastanza performanti, ma il migliore è il **RecogniserSigned.py**.

1. Risultato Ottenuto

L’obiettivo che ci eravamo prefissati di ottenere con questa tesina è stato raggiunto, in quanto abbiamo ottenuto un *assistente visivo virtuale* che in real-time è in grado di riconoscere tutti i gesti rappresentanti le lettere dell’alfabeto del Signuno.

Immagine che contiene testo, interni, guanti

Descrizione generata automaticamenteImmagine che contiene testo, persona

Descrizione generata automaticamenteImmagine che contiene testo, interni, maschio

Descrizione generata automaticamenteImmagine che contiene testo, persona, interni, schermo

Descrizione generata automaticamenteIn seguito, alcune immagini dimostrative rappresentanti il riconoscimento di qualche gesto del Signuno, usando il **Recogniser Signed**:

Ciò che potrebbe essere migliorato è il *grado di precisione del riconoscimento dei gesti*, in quanto non sempre il programma è in grado di riconoscerli correttamente. A volte, infatti, pur facendo il gesto correttamente, come riportato in figura 2.1, quest’ultimo non viene riconosciuto o scambiato per un altro, mentre altre volte pur non eseguendo il gesto nel modo delineato in figura 2.1, il programma riesce a riconoscerlo.

1. Condizioni di Luminosità

**ROBERTA**

1. Requisiti Hardware e Software

Caratteristiche hardware e software dei computer che abbiamo usato per lo sviluppo e test del progetto:

* CPU: Intel Core i7 8th Gen
* GPU: Nvidia GeForce GTX 1050
* Webcam: HD 720p
* Sistema operativo: Windows
* Python versione 3.9
* IDE: PyCharm
* Librerie: Mediapipe, OpenCV, matplotlib e altre più basiche